DOCUMENT RESUME

ED 272 528 TM 860 441

AUTHOR Wise, Lauress L.
TITLE Latent Trait Models for Partially Speeded Tests.
INSTITUTION American Inst. for Research, Washington, DC. Washington Research Center.
SPONS AGENCY Association of American Medical Colleges, Washington, D. C.
PUB DATE Apr 86
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Entrance Examinations; *Difficulty Level; Educational Research; *Error of Measurement; *Item Analysis; *Latent Trait Theory; *Mathematical Models; Predictive Measurement; Psychometrics; Scoring; Statistical Studies; *Test Items
IDENTIFIERS *Medical College Admission Test; Speededness (Tests)

ABSTRACT

A primary goal of this study was to determine the extent to which item difficulty was related to item position and, if a significant relationship was found, to suggest adjustments to predicted item difficulty that reflect differences in item position. Item response data from the Medical College Admission Test (MCAT) were analyzed. A data set was created containing statistical data from two uses of each of 9,099 items. Three models were investigated for describing the relationship between item position and the decrease in the adjusted percent passing: (1) a linear model; (2) a quadratic model; and (3) an exponential model. Results show that item statistics can be significantly related to item position even for tests that are not apparently speeded. Taken as a whole, the results suggest a more comprehensive model of item difficulty with the time spent on each item as a mediating variable. Further development of such a model should be based on data from computer-administered tests where response latency data can be collected and analyzed. Other areas for future research include empirical investigations of the relationship of item position effects to examinee performance level and to characteristics of preceding items in the form. (LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED272528

Latent Trait Models for Partially Speeded Tests

Lauress L. Wise
American Institutes for Research

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
L. L. Wise
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the 1986 Annual Convention of the American Educational Research Association, April 19, San Francisco

The research reported here was supported by the Association of American Medical Colleges. Statements and conclusions are those of the author and do not necessarily reflect the position or views of the Association.

TM 860 441

WASHINGTON RESEARCH CENTER
1055 THOMAS JEFFERSON STREET, NW
WASHINGTON, DC 20007
(202) 342-5000

INTRODUCTION

Item Response Theory (IRT) models are increasingly being employed in the precalibration, if not the actual scoring, of large-scale national tests. Precalibration of forms involves measures to assure that different forms of a test are as "parallel" as possible. The primary concern is that the error of measurement at each score level be as constant across forms as possible, so that an examinee has the same chance of getting scores slightly above or below his or her true performance level across all test forms. The second condition for parallelism, that each examinee should have the same expected score across alternative forms, is usually handled through post-administration equating procedures. (See, for example, Lord & Novick, 1968 for a rigorous explication of parallel forms.)

The error of measurement at a given "true-score" level is a function of the distribution of item difficulties and item discrimination levels. Even when traditional item statistics for item difficulty (the proportion passing the item) and discrimination (item-total biserial correlations) are used in form precalibration, IRT models can play an important role. Items included in each new form have almost always been previously calibrated based on statistics from a prior use either in a prior scored form or in an experimental use with a smaller sample of examinees. An important step in form precalibration is putting statistics from different prior administrations of items onto a common scale. This step is particularly important where testing at different times during the year results in significant variations in the performance levels of examinees at different testing sessions. IRT models seek to provide statistics that are constant across different populations of examinees except for a linear adjustment to equate item paramteters for differences in examinee performance mean and homogeneity levels.

Formulas for adjusting traditional item statistcs for differences in examinee performance distributions were developed in a prior study (Wise & Wilson, 1982). In that study, a two parameter model yielded significant improvements in prediction over a one parameter model as the degree of change in the proportion passing an item was significantly related to item biserials. Adding a third parameter to the model did not significantly improve the prediction of traditional item statistics.

The present study sought to identify other aspects of the prior administration of an item that might be used to improve prediction of the proportion passing the item in a new use. In particular, it had been suggested that item difficulty might be related to item position, even though there was no evidence that the tests in question were at all "speeded". (Generally, 98% to 99% of all examinees reach the last item of each test.) The primary goal of the present study, therefore, was to determine the extent to which item difficulty was related to item position and, if a significant relationship was found, to suggest adjustments to predicted item difficulty that reflect differences in item position.

METHODS

Item response data from the Medical College Admission Test (MCAT) were analyzed. A data set was created containing statistical data from two uses of each of 9,099 items. In analyzing the effects of item position, data from the use in which the item had the lowest item number was designated predictor data. Data from the use in which the item had the highest item position was designated "target" data. Where an item was used twice in the same position, the use with the larger sample size was designated as the predictor use. If these tests are at all speeded, the items should exibit a decrement in performance as they are placed later in a form. The goal of these analyses was to determine how much lower the proportion passing was when an item was used later in the test.

Data were taken from the four separately timed sections of the MCAT: Science Knowledge (125 items per form), Science Problems (66 items per form), Reading Skills (68 items per form), and Quantitative Skills (also 68 items per form). The primary design of these analyses was to predict the proportion passing the item in its "target" use from information on the proportion passing in the "predictor" use of the item and from information on the items position in the two uses. In addition, samples of examinees receiving each form varied in terms of their true performance levels (scaled in terms of the 15-point MCAT equated score scale) so estimates of the mean and standard deviation of examinee performance levels for each sample of examinees were included. Table 1 presents a statistical summary of the predictor and target information for each of these separate sections.

In the analyses that follow, the proportions passing each occurrence of each item were converted to probits using the inverse cummulative normal distribution function. Successive "error" terms were computed. The first was just the difference between the target probit value and the predictor probit value. The second error term was the difference between the target probit value and a predicted value based on information about the scale score means and standard deviations for the predictor and target samples and the item's biserial in the predictor use. This adjustment, developed in the prior study, was:

$$Xt = \{ Xp + Bp*(Mt-Mp)/Sp \} / SQRT\{ 1-Bp^{2}*(1-St/Sp)^{2} \}$$

where Xt is the predicted target probit, Xp and Bp are the difficulty probit and biserial from the predictor use of the item and Mt, Mp, St, and Sp are estimates of the overall performance score mean and standard deviations for the target and predictor use samples.

Table 1
Predictor and Target Sample Characteristics
For Each Subtest

Section	Variable	Predictor Mean	Predictor S.D.	Target Mean	Target S.D.
Science Kn.	Item Number	107.4	11.8	114.0	11.5
(N=964,	% Corr. Prbt	.11	.53	.06	.50
Phys only)	Sample Mean	8.08	.32	8.01	.39
	Sample S.D.	2.55	.096	2.54	.116
	Form Mean	28.1	1.8	28.4	1.7
	Form S.D.	8.8	.7	8.8	.6
Science Pr.	Item Number	32.4	19.2	46.8	19.8
(N=1589)	% Corr. Prbt	.22	.51	.19	.50
	Sample Mean	7.94	.31	7.95	.34
	Sample S.D.	2.48	.09	2.48	.12
	Form Mean	34.5	2.1	35.2	1.9
	Form S.D.	9.4	.6	9.4	.6
Reading Sk.	Item Number	40.4	21.8	51.9	18.9
(N=2228)	% Corr. Prbt	.74	.57	.67	.54
	Sample Mean	7.74	.46	7.78	.38
	Sample S.D.	2.49	.21	7.30	.53
	Form Mean	46.1	1.6	45.8	1.5
	Form S.D.	7.3	.6	7.3	.5
Quant. Sk.	Item Number	39.9	21.6	52.8	18.0
(N=2154)	% Corr. Prbt	.36	.62	.33	.60
	Sample Mean	7.58	.39	7.60	.40
	Sample S.D.	2.46	.13	2.49	.13
	Form Mean	42.2	1.6	42.2	1.3
	Form S.D.	8.2	.7	8.3	.7

NOTE: Because Performance Mean and Form Difficulty information is maintained separately for each of the separate discipline scores within science knowledge, only physics items were included in this table to avoid mixing noncomparable statistics. The physics items come last and are at most risk of significant item position effects. The Form Mean and S.D. statistics are the expected raw score means and S.D.s for an examinee population with a "scaled score" mean of 8.0 and standard deviation of 2.5.

RESULTS

Table 2 shows the correlation of the unadjusted and adjusted error terms with various statistics. As can be seen, the adjustment for examinee sample differences is generally successful in eliminating differences in item performance associated with differences in these sample characteristics. There is also a significant relationship between item performance and item position. This significance is generally increased when variation due to examinee sample differences is removed.

Table 2
Correlation of Prediction Errors
with Position and Examinee and Form Characteristics

	Unadjusted Error				Adjusted Error			
Statistic	SK	SP	R	Q	SK	SP	R	Q
Samp Mean Change	.62	.53	.59	.54	.08	.04	.02	.14
Samp S.D. Ratio	.39	.30	-.52	-.03	.08	-.00	.02	.04
Item Pos. Change	-.35	-.35	-.22	-.14	-.46	-.34	-.43	-.27

Given a strongly significant relationship between item position and item difficulty, the next step was to model the relationship between item position and the decrease in the adjusted percent passing. Three models were investigated: (1) a linear model, (2) a quadratic model, and (3) an exponential model. The data for assessing item position effects were all relative; the performance of an item at one position was compared to the performance at other positions. Since there was no absolute measure of position effect, an arbitrary zero point was set at item position 1, where no position effect was assumed. The three models were thus as follows:

(1) $TD = APD - a*(Tpos - Ppos)$,

(2) $TD = APD - a*(Tpos - Ppos) - b*(Tpos^2 - Ppos^2)$, and

(3) $TD = APD - EXP(a*Tpos+b) + EXP(a*Ppos+b)$;

where TD is the Target difficulty (% correct probit), APD is the adjusted Predictor difficulty, Tpos and Ppos are the target and predictor item positions, and a and b are model constants to be estimated.

Table 3 shows the parameter estimates that resulted from fitting each of the three models to the available data for each of the four MCAT sections using the SAS procedure NLIN. All of the subtests showed a highly significant linear trend. The Reading subtest showed the greatest per-item decrease in performance levels. The two parameter models are somewhat more difficult to interpret as the parameter estimates are highly correlated. Table 4 provides a summary of the significance of each model in reducing prediction errors. For each of the three models, the Error Mean Square (EMS) is shown along with an F statistic testing the significance of the model in explaining target item performance values after already

controlling for adjusted predictor item perfomance values. In addition, F statistics testing the significance of the improvement of the quadratic and exponential models over the linear model are given.

Table 3
Item Position Parameter Estimates
(and asymptotic standard errors)
For each Model and Subtest

		Model				
		Linear	Quadratic		Exponential	
Subtest	# of Itms	a	a	b	a	b
Science Kn	2837	-.0014	-.0020	.000026	.055	-9.47
		(.0002)	(.0004)	(.000008)	(.009)	(1.12)
Science Pr.	1416	-.0024	-.0002	.000032	.032	-3.97
		(.0001)	(.0007)	(.000009)	(.010)	(.76)
Reading Sk.	1946	-.0050	-.0011	.000072	.049	-4.46
		(.0002)	(.0010)	(.000013)	(.008)	(.55)
Quant. Sk.	1905	-.0021	-.0033	.000070	.070	-6.70
		(.0002)	(.0008)	(.000010)	(.017)	(1.15)

Table 4
Fit of Alternative Models of Item Position Effects

	MODEL						F of Improvement over Linear	
	1 LINEAR		2 QUAD.		3 EXP.			
Section	EMS	F	EMS	F	EMS	F	Quad.	Exp.
Science Kn.	.0123	83.8	.0118	108.5	.0115	142.9	110.7	119.3
Science Pr.	.0134	310.9	.0133	164.5	.0133	164.4	11.6	11.6
Reading Sk.	.0300	638.8	.0292	354.2	.0292	358.9	54.3	61.2
Quant. Sk.	.0263	147.2	.0256	101.0	.0257	98.0	53.1	45.5

All of the models explain significant variance in the item difficulty residuals. The two parameter models (quadratic and exponential) explain significantly more of the residual variance than the one parameter, linear model in all cases. While there is not a clear difference between the quadratic and exponential models, there might be some reason to prefer the exponential model based on the results shown in Table 4.

One further set of analyses was run to identify other characteristics of the items, test forms, or examinees that might also be related to changes in item performance. A third error term was computed using the parameter estimates for the exponential model of item position effects to adjust predicted difficulties for changes in item position. Table 5 shows correlations of selected item, form, and examinee characteristics with each of the three error terms computed in the study.

Table 5
Correlation of Item, Form, and Examinee Statistics
with Successive Residual Error Terms
(N=6935 items)

Statistic	Unadj. Error	Adj. for Examinee Diffs.	Also Adj. for Item Pos.
Pop. Mean Change	.56	.05	.13
Pop. S.D. Ratio	-.23	.02	-.06
Item Pos. Change	-.21	-.34	-.00
Target Form Mean	-.05	-.10	-.03
Form Mean Change	.03	.08	.07
Target Form S.D	.01	.03	-.04
Form S.D. Ratio	-.13	-.18	-.21

Two of the results shown in Table 5 are worthy of note. First, the correlation with the mean change in the population performance level increased from .05 to .13 after the item position correction was introduced. The item position correction appeared to overcorrect (subtract too much from) predicted passing rates for higher levels of target population examinee performance. That the tests appear less "speeded" for higher performing groups of examinees seems entirely rational. A further attempt was made to investigate this finding. Items were divided into four roughly equal groups on the basis of the mean performance level of the target groups. Separate regression slopes were estimated for each item group. The results are shown in Table 6. For two of the four subtests, there were no real differences in the regression slopes for the four different groups. For the other two subtests, the slopes were noticeably steeper for the examinee groups with lower mean performance levels.

Table 6
Regression Slopes for Predicting Proportion Passing from Position For Each Subtest and Examinee Performance Mean Group

Subtest	Examinee Group Performance Means: Lowest Quartile	2nd Quartile	3rd Quartile	Highest Quartile
Science Kn.	-.0015	-.0013	-.0011	-.0019
Science Pr.	-.0021	-.0021	-.0028	-.0028
Reading Sk.	-.0066	-.0078	-.0049	-.0035
Quant. Sk.	-.0052	-.0047	-.0019	-.0028

In interpreting Table 6, it should be noted that data on individuals were not available for these analyses. While it was possible to separate items on the basis of group performance means, the variance in these means was small in comparison to the within-group variation. Clearer differences in test "speededness" might have resulted if examinees had been separated into non-overlapping performance strata.

A second result of note in Table 5 is the significant correlation of the Form S.D. ratio index with residual error. The Form S.D. statistic is the expected raw score standard deviation for a particular reference population (Spring 1977 examinees). The ratio will be high when the expected raw score variance is greater for the target examinee group than for the predictor examinee group. The expected raw score variance is higher if the item biserials are higher and/or if the item difficulties are more homogeneous (given a mean difficulty near .5). The negative correlation in Table 5 thus suggests that item passing rates drop faster when the item biserials are higher or when the item difficulties are more homogeneous. Unfortunately, more detailed information on the characteristics of items comprising each of the forms was not available for these analyses.

DISCUSSION

The results of this study show that item statistics can be significantly related to item position even for tests that are not apparently speeded. There is evidence that the relationship between item position and decrement in the proportion passing the item is curvilinear with increasing decrements in examinee performance as the item is placed further back in the form. Even with the very large item sample sizes, however, it was difficult to determine a more precise functional relationship empirically.

Taken as a whole, the results suggest a more comprehensive model of item difficulty with the time spent on each item as a mediating variable. We would need to specify time spent on an item as a function of item difficulty, examinee performance level, and time remaining and then model the probability of passing an item as a joint function of examinee ability level, item characteristics, and time spent on the item. Further development of such a model should be based on data from computer-administered tests where response latency data can be collected and analyzed.

Other areas for further research include empirical investigations of the relationship of item position effects to examinee performance level and to characteristics of preceding items in the form. The relationship of the degree of item position effect to other item characteristics such as the amount and level of required reading, the use of graphics, and amount of computation required might also be examined.

From a practical perspective, it may be sufficient to derive item position effect curves empirically for a given test and make use of this information in the precalibration of forms. Alternatively, techniques such as section pre-equating that leave item position constant also avoid mis-calibration due to changes in item position.

REFERENCES

Lord, F., & Novick, M. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Wise, L., & Wilson. S. (1982). "Test item calibration." Paper presented at the annual convention of the American Educational Research Association, New York.